Structural Sparsity in Multiple Measurements

F. Boßmann, S. Krause-Solberg, J. Maly, and N. Sissouno

***Abstract*—We propose a novel sparsity model for distributed compressed sensing in the multiple measurement vectors (MMV) setting. Our model extends the concept of row-sparsity to allow more general types of structured sparsity arising in a variety of applications like, e.g., seismic exploration and non-destructive testing. To reconstruct structured data from observed measurements, we derive a non-convex but well-conditioned LASSO-type functional. By exploiting the convex-concave geometry of the functional, we design a projected gradient descent algorithm and show its effectiveness in extensive numerical simulations, both on toy and real data.**

***Index Terms*—Distributed compressed sensing, multiple measurements, sparse approximation, structured sparsity, non-convex LASSO.**

I. Introduction

STARTING with the seminal works [8], [9], [16] a rich theory on signal reconstruction from seemingly incomplete information has evolved under the name *compressed sensing* in the past two decades, cf. [17] and references therein.

The core idea is to use the intrinsic structure of a high-dimensional signal $\mathbf{x} \in \mathbb{R}^N$ to allow reconstruction from $m \ll N$ linear measurements

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times N}$ models the measurement process and $\mathbf{y} \in \mathbb{R}^m$ is called the *measurement vector*. In what follows we will call the columns $\mathbf{a}_j$ of $\mathbf{A}$ *atoms*. One particular instance of intrinsic structure is *sparsity:* the signal $\mathbf{x}$ is called s-sparse if $|\mathrm{supp}(\mathbf{x})| \leq s$, where $\mathrm{supp}(\mathbf{x}) = \{i : x_i \neq 0\}$ denotes the *support* of $\mathbf{x}$. We will call an atom $\mathbf{a}_j$ *activated* if $j \in \mathrm{supp}(\mathbf{x})$. If $\mathbf{A}$ is well-designed, $m \approx s\log(\frac{N}{s})$ measurements suffice to guarantee stable and robust recovery of all s-sparse $\mathbf{x}$ from $\mathbf{y}$ by polynomial time algorithms [17]. This suggests that the number of necessary measurements mainly depends on the intrinsic information encoded in $\mathbf{x}$.

Despite its simplicity the model in (1) encompasses many real-world measurement set-ups. For instance, in seismic exploration, a key challenge is to reconstruct earth layers from few linear measurements [5], [6]. In this case, the vector $\mathbf{x}$ is a discretized vertical slice through the ground where each entry represents the seismic reflectivity. To reconstruct the earth layers, a synthetic seismic impulse is produced and its reflections are measured at different positions on the surface. This measurement process can be modeled by a convolution of $\mathbf{x}$ with the seismic impulse such that $\mathbf{A}$ is the corresponding convolution matrix. Since the reflectivity is low whenever the material is mostly homogeneous and high at material boundaries, the vector $\mathbf{x}$ can be assumed to be sparse and its non-zero entries indicate the earth layer boundaries. A similar model applies to ultrasonic non-destructive testing where an ultrasonic impulse is sent into an object and defects inside the material are reconstructed from the reflections of this impulse [7]. Other possible applications are face and speech recognition [21], [36], magnetic resonance imaging [25], or computer tomography [30]. For an overview also see [27], [31] and the references therein.

In all applications mentioned above, we can assume structure not only in one direction of space but in multiple dimensions, meaning that measurements at different locations/times $t_1 < \ldots < t_L$ correspond to different ground-truth signals $\mathbf{x}_1, \ldots, \mathbf{x}_L \in \mathbb{R}^N$ whose support structure is related. When thinking of waves travelling through the ground, the positions of non-zero entries of consecutive $\mathbf{x}_l$ can only differ up to a certain number determined by properties of the surrounding material and fineness of the discretization.

If several signals $\mathbf{x}_1, \ldots, \mathbf{x}_L \in \mathbb{R}^N$ are measured by the same process $\mathbf{A}$, the model in (1) becomes

$$\mathbf{Y} = \mathbf{A}\mathbf{X}, \tag{2}$$

for $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_L) \in \mathbb{R}^{N\times L}$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_L) \in \mathbb{R}^{m\times L}$. In this setting, also known as *multiple measurement vectors (MMV)*, the necessary number of measurements may be reduced by exploiting joint structure in $\mathbf{X}$, for instance, row sparsity (all $\mathbf{x}_l$ share a common support). This has already been done in applications like MRI [37] and MIMO communications [28]. In the case of row-sparsity, reconstruction is usually performed via

$$\min_{\mathbf{X}\in\mathbb{R}^{N\times L}} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_F^2 + \lambda\|\mathbf{X}\|_{\text{row}-0} \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius-norm, $\|\cdot\|_{\text{row-0}}$ denotes by abuse of notation the number of non-zero rows, and λ is a

Manuscript received March 2, 2021; revised August 13, 2021; accepted December 12, 2021. Date of publication December 23, 2021; date of current version January 12, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sheetal Kalyani. The work of F. Boßmann was supported by the National Science Foundation of China under Project 42004109, (Seismic data interpolation using wave model decomposition). The work of S. Krause-Solberg was supported by Helmholtz Imaging, a platform of the Helmholtz Incubator on Information and Data Science. The work of J. Maly was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the project CoCoMIMO funded within the priority program SPP 1798 *Compressed Sensing in Information Processing* (COSIP). *(Corresponding author: F. Boßmann.)*

F. Boßmann is with the Harbin Institute of Technology, Department of Mathematics, Harbin 150001, China (e-mail: f.bossmann@hit.edu.cn).

S. Krause-Solberg is with the Helmholtz Imaging, Deutsches Elektronen-Synchrotron DESY, D-22607 Hamburg, Germany (e-mail: sara.krause-solberg@desy.de).

J. Maly is with the KU Eichstätt, D-85072 Eichstätt, Germany (e-mail: johannes.maly@ku.de).

N. Sissouno is with the Technical University of Munich, Faculty of Mathematics, D-85748 Garching, Germany (e-mail: sissouno@ma.tum.de).

Digital Object Identifier 10.1109/TSP.2021.3137599

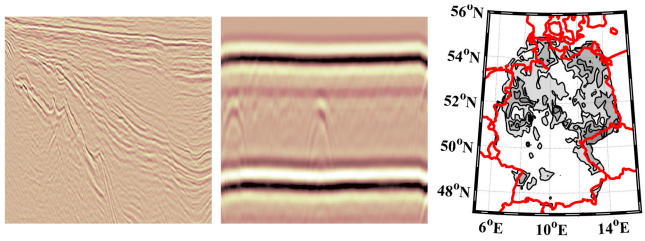


Fig. 1. Data from different applications exhibiting clear structure but being neither row- nor block-sparse: Seismic exploration (left), ultrasonic non-destructive testing (middle), and meteorology (right).

tunable parameter. Although (3) is NP-hard in general [17], solutions can be well approximated via greedy algorithms or convex relaxation.

However, row-sparsity and other established structural models (column sparsity, block sparsity) make restrictive assumptions on the concrete structure of $\mathbf{X}$. As a simple thought experiment, consider $\mathbf{X}$ to be the identity matrix. Then $\mathbf{X}$ is neither row- nor block-sparse but clearly exhibits a simple structure, namely a diagonal line. In fact, the established models are too restrictive for many applications. For example, the support and thus the structure may change over time as it is the case in real-time dynamic MRI [34], dynamic PET [18], and wireless communication [23]. In machine learning as well as in quantile and logistic regression modeling more sophisticated structure models may be required [19], [22]. In this work, we consider the three applications seismic exploration, ultrasonic non-destructive testing, and meteorology. Here, the data is gathered at different locations where the support might change over the spatial dimension. For instance, the above mentioned earth layers in seismic exploration do not follow straight horizontal lines and thus are not even close to row sparse. Material defects that have to be reconstructed in non-destructive testing can have many forms, most of which are neither row nor block sparse. In Fig. 1 we show some exemplary measurements that do not fit into established structural models.

Moreover, in some applications there does not exist a left-to-right order of the measurement vectors $\mathbf{y}_l$ (for instance, when the measurements are taken from scattered locations) and also the order of atoms $\mathbf{a}_j$ may not be clear. In this case, we want to reconstruct the solution independent of column permutations in $\mathbf{A}$ and $\mathbf{Y}$. If $\mathbf{X}$ is the structural sparse solution of $\mathbf{AX} = \mathbf{Y}$, then $\tilde{\mathbf{X}} = \mathbf{P_Y X P_A}$ should be the structural sparse solution of $(\mathbf{P_A A})\tilde{\mathbf{X}} = \mathbf{P_Y Y}$, where $\mathbf{P_A}$ and $\mathbf{P_Y}$ are permutation matrices. This excludes all order-dependent approaches to define structural sparsity since such definitions are not invariant under permutations.

A. Contribution

In this work, we introduce a sparsity model that can capture a wide range of practically relevant structures, comes with efficient optimization, and allows to learn the structures in an intuitive way from only the measurements and additional knowledge of the concrete application. To be more precise, our model is a special case of group sparsity for multiple measurement recovery problems and encompasses established concepts like row- and block-sparsity. The novel ingredient is that the structural support constraints are encoded in a matrix $\mathbf{C}$ which allows efficient processing. We introduce a non-convex regularizer enforcing the structures encoded in $\mathbf{C}$ and discuss possible relaxations of the regularizer. Based on theoretical insights into the optimization landscape of our regularizer, we suggest a projected gradient descent to minimize the related LASSO-type formulation. Finally, we provide a simple heuristic to determine $\mathbf{C}$ for concrete applications under sole knowledge of the measurement process and measurements. We validate efficacy of both the parameter heuristic and the regularizer in extensive numerical simulations on real data.

B. Related Work

There exist several approaches to solve (2) by assuming different sparsity models for $\mathbf{X}$. Most of them adapt methods that were originally designed to solve (1). In [13] an extension of the algorithms Matching Pursuit and the FOCal Underdetermined System Solver (FOCUSS) are presented. Bayesian methods are considered in [41], [46]. In [32], [33] the authors introduce greedy pursuits and convex relaxations for the MMV problem. Theoretical results have been shown, e.g., in [12]. All these methods enforce row-sparsity in the reconstructed solution. In [4] two joint sparsity models (JSM) for compressed sensing are introduced. JSM-1 considers solutions where all columns can be written as the sum of a common sparse component that is equal for each column and another unique sparse vector. This is equivalent to assuming $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$ where $\mathbf{X}_1$ is row-sparse and $\mathbf{X}_2$ is sparse (without any correlation between different columns). JSM-2 is a slight relaxation of row-sparsity and allows small support changes over the columns. Yet another approach is presented in [24], [43]: correlated measurements are assumed to have sparse approximations that are close in the Euclidean distance. This idea is related to dynamic compressed sensing [2], [45] where neighboring columns are assumed to have similar support. In both cases, the support is allowed to change slowly over different data vectors. Nevertheless, the above methods share quite restrictive support assumptions based on geometrical features and hence cannot reconstruct simple features in the solution that do not match those strict assumptions.

A more general approach is the so-called group sparsity model [3], [20] in which a set of groups $\mathcal{G}$ is defined whose elements $G \in \mathcal{G}$ encode support sets. The matrix $\mathbf{X}$ is called s-group-sparse if $\operatorname{supp}(\mathbf{X})$ is a subset of a union of at most s groups in $\mathcal{G}$. Whereas this model is able to encode all possible structural constraints on $\mathbf{X}$, its generality comes with a price. First, straight-forward adaption of compressed sensing algorithms is only possible under knowledge of the set $\mathcal{G}$, cf. [3], [20] and subsequent literature. Second, even if $\mathcal{G}$ is known, its cardinality might grow exponentially in the ambient dimension which considerably increases the computational complexity of established procedures. Third, if $\mathcal{G}$ is unknown, learning algorithms have to either rely on clustering of entries [38], [42] leading to block-sparse-like models, or on available training data $(\mathbf{X}_i, \mathbf{Y}_i)$, for $i \in [n]$, from which $\mathcal{G}$ can be learned [26]. The latter work suggests an alternating approach to learn both signal and

structure without initial data; nevertheless, it does not provide as simple means to incorporate expert knowledge on concrete applications as our heuristic for determining the structure encoding matrix $\mathbf{C}$. Let us finally mention that, for general $\mathcal{G}$, convex regularizers can only be computed theoretically, cf. the concept of atomic norms in [11].

Remark 1: In this work we concentrate on understanding and representing the intrinsic sparsity structure of $\mathbf{X}$ for a fixed measurement process/dictionary $\mathbf{A}$. Other lines of work discuss how to learn a proper dictionary for $\mathbf{X}$ (such that $\mathbf{X}$ becomes sparse in a classical sense) or how to adapt an existing one by slight perturbation to improve reconstruction performance [1], [44]. They, however, do not allow to represent general structural dependencies between active entries of $\mathbf{X}$. Combining those approaches with our generalized sparsity model is an interesting topic for future work.

C. Notation and Outline

We denote matrices by bold capital letters, vectors by bold lowercase letters, and scalars by regular letters. The only exceptions are vectors that we get by the vectorization $\vec{\mathbf{Z}} := \mathrm{vec}(\mathbf{Z})$ of matrices $\mathbf{Z}$. The inversion (reshape) of the vectorization is denoted by vec^{-1}. As already mentioned in the introduction, the columns of a matrix $\mathbf{Z} \in \mathbb{R}^{m\times n}$ are denoted by $\mathbf{z}_l$ for $l \in [n]$, where $[n] := \{1, \dots, n\}$ is used to abbreviate index sets. The identity matrix and the matrix of ones are written as $\mathbf{Id}$ and $\mathbb{1}$, respectively. For the set of non-negative real numbers we use the notation $\mathbb{R}_+$.

We denote the support matrix of $\mathbf{Z}$ by $\mathbf{Z}_{01}$, i.e., $\mathbf{Z}_{01} \in \{0, 1\}^{m\times n}$ is the matrix with entries $|\mathrm{sign}(Z_{j,l})|$ for $j \in [m]$ and $l \in [n]$. If applied to matrices or vectors, the sign-function as well as the absolute value $|\cdot|$ act component-wise.

Besides the standard matrix multiplication we use the Kronecker product $\otimes$ defined by $\mathbf{A} \otimes \mathbf{B} := (A_{j,l}\mathbf{B})_{j,l} \in \mathbb{R}^{mp\times nq}$, for matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{p\times q}$. The Frobenius norm of a matrix $\mathbf{Z}$ is $\|\mathbf{Z}\|_F$, while $\|\mathbf{z}\|_2$ denotes the Euclidian norm of a vector $\mathbf{z}$.

The outline of the paper is as follows. In Section II, we introduce a general model for structural sparsity and discuss possible reconstruction approaches of such structured signals. In particular, we derive a relaxed, still non-convex functional, whose minimizers provide good approximations to $\mathbf{X}$, and explore the specific geometry of the functional. Building upon these insights, we describe in Section III a projected gradient descent procedure to efficiently solve the program. Finally, in Section IV, we empirically validate our model on toy scenarios and real (seismic/ultrasonic/meteorological) data.

II. Structural Sparsity

We begin by deriving a general model for structural sparsity. After discussing its relation to established structural models like row- or column-sparsity, we provide a corresponding NP-hard optimization problem to reconstruct structured signals from compressive measurements. To solve the in general intractable problem, we suggest a non-convex relaxation of particular convex-concave shape.

A. The Basic Model

In order to develop a notion of structural sparsity that is capable of describing signals like the ones in Fig. 1, we first have to understand the underlying abstract idea of row-sparsity and related concepts. A matrix $\mathbf{X}$ is s-row-sparse if it has at most s non-zero rows, i.e., if there are up to s matrices $\mathbf{X}_k$ with exactly one non-zero row such that $\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_s$. We could say that the matrices $\mathbf{X}_k$ describe the *elementary structures* of row-sparsity. By changing the elementary structures, one obviously recovers various established concepts like sparsity ($\mathbf{X}_k$ are matrices with exactly one non-zero entry) and block-sparsity ($\mathbf{X}_k$ are matrices with exactly one non-zero block). This elementary idea is the corner stone of group sparsity [3], [20].

Building upon the same intuition, we wish to describe elementary structures in a practical way that lends itself to efficient computation. To this end, given two non-zero entries $X_{j,l}, X_{j',l'}$ of a structured signal matrix $\mathbf{X} \in \mathbb{R}^{N\times L}$, let $C_{(j,l),(j',l')} \in \{0, 1\}$ indicate whether the two entries can belong to a single elementary structure of $\mathbf{X}$. To be more precise, $C_{(j,l),(j',l')} = 0$ if they can belong to the same structure, i.e., there exists (at least) one elementary structure $\mathbf{X}_k$ whose entries (j, l) and (j', l') are non-zero. If such a structure does not exist, we set $C_{(j,l),(j',l')} = 1$. Then,

$$\sum_{\substack{j,j',l,l':\\ X_{j,l},X_{j',l'}\neq 0}} C_{(j,l),(j',l')} = 0 \tag{4}$$

whenever $\mathbf{X}$ itself is an elementary structure. Moreover, the value of (4) increases the more $\mathbf{X}$ differs from an elementary structure. Let us clarify this by a simple example: the choice $C_{(j,l),(j',l')} = 0$ for $j = j'$ and 1 otherwise would characterize the basic units of row-sparsity as (4) is 0 if and only if $\mathbf{X}$ has at most one non-zero row. Using the vectorization of the support matrix $\mathbf{X}_{01} \in \{0, 1\}^{N\times L}$ of $\mathbf{X} \in \mathbb{R}^{N\times L}$ we can rewrite (4) as

$$\vec{\mathbf{X}}_{01}^T \mathbf{C} \vec{\mathbf{X}}_{01} = 0, \tag{5}$$

where $\mathbf{C} \in \mathbb{R}^{NL\times NL}$ has entries $C_{(j,l),(j',l')}$, for $j, j' \in [N]$ and $l, l' \in [L]$. We may now define the set of $\mathbf{C}$-structured s-sparse signals

$$\mathbb{S}^s_{\mathbf{C}} = \left\{ \mathbf{Z} \in \mathbb{R}^{N\times L} : \begin{array}{c} \mathbf{Z} = \sum_{k=1}^s \mathbf{Z}_k, \\ (\vec{\mathbf{Z}}_k)_{01}^T \mathbf{C} (\vec{\mathbf{Z}}_k)_{01} = 0, \ \forall k \in [s] \end{array} \right\}. \tag{6}$$

Remark 2: The model defined in (6) is quite general and covers several well-known special cases (in addition to row-sparsity mentioned above). Choosing $\mathbf{C} = \mathbb{1} - \mathbf{Id}$, the set $\mathcal{S}^s_{\mathbf{C}}$ describes the set of s-sparse vectors. Choosing $\mathbf{C}$ such that

$$C_{(j,l)(j',l')} = \begin{cases} 0 & |j - j'| \le a \text{ and } |l - l'| \le b, \\ 1 & \text{else,} \end{cases}$$

we recover the set of block-sparse matrices with blocks of size $a \times b$ (cf. [3]).

Note that the diagonal entries of $\mathbf{C}$ are 0 independent of the concrete model as $\mathbf{C}$ shall only characterize relations between different entries of $\mathbf{X}$.

Building upon (6) we can define the $\mathbf{C}$-structured ℓ_0-norm

$$\|\mathbf{Z}\|_{\mathbf{C},0} = \min\{s \ge 0 : \mathbf{Z} \in \mathbb{S}^s_{\mathbf{C}}\}, \tag{7}$$

which is actually not a norm but abuse of notation. Minimizing the $\ell_{\mathbf{C},0}$-norm constrained to correct measurements, i.e.,

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times L}} \|\mathbf{Z}\|_{\mathbf{C},0}, \quad \text{subject to } \mathbf{AZ}=\mathbf{Y}, \tag{8}$$

then extends ℓ_0-minimization to the $\mathbf{C}$-structured s-sparse case. The program in (8) inherits NP-hardness from classical sparse recovery such that it is undesirable to solve (8) directly. In fact, even computing (7) is NP-hard in general. Note, however, that (5) itself might suffice as regularizer since its magnitude increases if the number of elementary structures in $\mathbf{X}$ increases. This handwavy argument is substantiated by the observation that, under mild assumptions on the elementary structures encoded in $\mathbf{C}$, there is an equivalence relation between (5) and (7) as stated in the following proposition.

Proposition 3: Assume that, for $s_{\text{col}} \in [N]$, the sparsity model characterized by $\mathbf{C}$ satisfies

$$\|\mathbf{X}\|_{\mathbf{C},0} = 1 \Rightarrow \|\mathbf{x}_l\|_0 \leq s_{\text{col}}, \ \forall l \in [L]. \tag{9}$$

Then, for $\mathbf{X} \neq \mathbf{0}$, we have

$$\frac{1}{2}\|\mathbf{X}\|_{\mathbf{C},0}^2 \leq \vec{\mathbf{X}}_{01}^T \mathbf{C}\vec{\mathbf{X}}_{01} + 1 \leq s_{\text{col}}^2 L^2 \|\mathbf{X}\|_{\mathbf{C},0}^2.$$

Proof: For $\|\mathbf{X}\|_{\mathbf{C},0} = 1$, one has that $\vec{\mathbf{X}}_{01}^T \mathbf{C}\vec{\mathbf{X}}_{01} = 0$ by (6). Let us now assume $\|\mathbf{X}\|_{\mathbf{C},0} = s \geq 2$. Then we can write $\mathbf{X}_{01} = \sum_{k=1}^{s} \mathbf{X}_k$ where each $\mathbf{X}_k$ contains only ones and zeros. We hence get

$$\vec{\mathbf{X}}_{01}^T \mathbf{C}\vec{\mathbf{X}}_{01} = \left(\sum_{k=1}^{s} \vec{\mathbf{X}}_k\right) \mathbf{C} \left(\sum_{k=1}^{s} \vec{\mathbf{X}}_k\right) = \sum_{j\neq k} \vec{\mathbf{X}}_k^{\,T} \mathbf{C}\vec{\mathbf{X}}_j,$$

since $\vec{\mathbf{X}}_k^{\,T} \mathbf{C}\vec{\mathbf{X}}_k = 0$, for all $k \in [s]$. By assumption, the matrices $\mathbf{X}_k$ have at most $s_{\text{col}}L$ non-zero entries such that

$$1 \leq \vec{\mathbf{X}}_k^{\,T} \mathbf{C}\vec{\mathbf{X}}_j \leq s_{\text{col}}^2 L^2, \tag{10}$$

where the lower bound is a consequence of the minimal decomposition required in (7). We conclude that

$$s(s-1) \leq \sum_{j\neq k} \vec{\mathbf{X}}_k^{\,T} \mathbf{C}\vec{\mathbf{X}}_j \leq s_{\text{col}}^2 L^2 s(s-1),$$

which yields the claim.

Remark 4: Assumption (9) requires the elementary structures described by $\mathbf{C}$ to have s_{col}-sparse columns which holds for $s_{\text{col}} = 1$ when working with generalizations of row-sparsity. This can be easily seen in Fig. 1. Let us emphasize that the construction of $\mathbf{C}$ presented in Appendix A satisfies (9) if the appearing parameters α and β are suitably chosen.

If we have information on the structure of $\mathbf{X}$ in form of $\mathbf{C}$ and the measurements $\mathbf{A}$ are injective when restricted to $\mathbb{S}_{\mathbf{C}}^s$, Proposition 3 suggests to approximate $\mathbf{X}$ from $\mathbf{Y}$ by solving

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times L}} \vec{\mathbf{Z}}_{01}^T \mathbf{C}\vec{\mathbf{Z}}_{01}, \quad \text{subject to } \mathbf{AZ}=\mathbf{Y}. \tag{11}$$

This raises two questions: how can one obtain a suitable structure matrix $\mathbf{C}$ in general and how can the NP-hard optimization in (11) be solved? We refer the interested reader to Appendix A for a simple heuristic to construct $\mathbf{C}$ from $\mathbf{A}$ and $\mathbf{Y}$, and now address the second question.

B. Ways of Relaxation

The program in (11) poses two difficulties. First, the matrix $\mathbf{C}$ is not necessarily positive semi-definite and, second, the support vector $\vec{\mathbf{Z}}_{01}$ depends in a non-continuous way on $\mathbf{Z}$. To circumvent the latter, we define the regularizer

$$\mathcal{R}_{\mathbf{C}} : \mathbb{R}^{N\times L} \to \mathbb{R}, \quad \mathcal{R}_{\mathbf{C}}(\mathbf{Z}) = \vec{\mathbf{Z}}^T \mathbf{C}\vec{\mathbf{Z}},$$

and rewrite (11) as the binary program

$$\min_{\substack{\tilde{\mathbf{Z}}\in\{0,1\}^{N\times L},\\ \text{sign}(|\mathbf{Z}|)=\tilde{\mathbf{Z}},\\ \mathbf{AZ}=\mathbf{Y}}} \mathcal{R}_{\mathbf{C}}(\tilde{\mathbf{Z}}), \tag{12}$$

where sign as well as $|\cdot|$ act component-wise on matrices. As (12) illustrates, the non-continuous/non-convex dependence of $\vec{\mathbf{Z}}_{01}$ on $\mathbf{Z}$ lies in the relation between $\mathbf{Z}$ and the auxiliary variable $\tilde{\mathbf{Z}}$. Replacing the sign-function by identity (interpreting identity as a convex relaxation of sign) leads to

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times L}} \mathcal{R}_{\mathbf{C}}(|\mathbf{Z}|), \quad \text{subject to } \mathbf{AZ}=\mathbf{Y}. \tag{13}$$

Since (13) does not incorporate noise on the measurements, we replace it with the more robust formulation

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times L}} \|\mathbf{AZ}-\mathbf{Y}\|_F^2 + \lambda \mathcal{R}_{\mathbf{C}}(|\mathbf{Z}|), \tag{14}$$

where $\lambda > 0$ is a regularization parameter. It is well known from related programs that solutions of (14) solve a robust version of (13) while the magnitude of λ balances robustness and accuracy of the reconstruction. We detail this in the following lemma. The proof is similar to [17, Proposition 3.2] and thus omitted.

Lemma 5: If $\mathbf{X}_\lambda$ minimizes (14) with parameter $\lambda > 0$, then $\mathbf{X}_\lambda$ minimizes

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times L}} \mathcal{R}_{\mathbf{C}}(|\mathbf{Z}|), \quad \text{subject to } \|\mathbf{AZ}-\mathbf{Y}\|_F \leq \eta_\lambda,$$

where $\eta_\lambda = \|\mathbf{AX}_\lambda - \mathbf{Y}\|_F$.

Although (14) is a non-convex problem, the regularizer $\mathcal{R}_{\mathbf{C}}$ exhibits some beneficial geometrical properties.

Proposition 6: For any $\mathbf{X}, \mathbf{D} \in \mathbb{R}^{N\times L}$ the following holds. Let $\mathbf{S} = \text{sign}(\mathbf{X}) \in \mathbb{R}^{N\times L}$ and define $\mathbb{R}_{\mathbf{S}}^{N\times L}$ as the orthant of $\mathbb{R}^{N\times L}$ in which $\mathbf{X}$ lies. Then,

$$f : \left\{t \in \mathbb{R} : \mathbf{X} + t\mathbf{D} \in \mathbb{R}_{\mathbf{S}}^{N\times L}\right\} \to \mathbb{R},$$
$$f(t) = \mathcal{R}_{\mathbf{C}}(|\mathbf{X} + t\mathbf{D}|) \tag{15}$$

is a convex or concave function. In particular, if $\mathbf{D} \in \mathbb{R}_{\mathbf{S}}^{N\times L}$ or $\mathbf{D} \in -\mathbb{R}_{\mathbf{S}}^{N\times L}$, then the function in (15) is convex.

Proof: Recall that all entries of $\mathbf{C}$ are non-negative and that $\mathbf{C}$ is symmetric by definition. Obviously,

$$g(t) = \mathcal{R}_{\mathbf{C}}(\mathbf{X} + t\mathbf{D}) = t^2 \vec{\mathbf{D}}^T \mathbf{C}\vec{\mathbf{D}} + 2t\vec{\mathbf{X}}^T \mathbf{C}\vec{\mathbf{D}} + \vec{\mathbf{X}}^T \mathbf{C}\vec{\mathbf{X}}$$

is a quadratic functional and thus either convex or concave. If $\mathbf{D} \in \pm\mathbb{R}_+^{N\times L}$, the functional g is convex as $\vec{\mathbf{D}}^T \mathbf{C}\vec{\mathbf{D}} \geq 0$. By restricting f such that $\mathbf{X} + t\vec{\mathbf{D}} \in \mathbb{R}_{\mathbf{S}}^{N\times L}$, we get that

$$f(t) = \mathcal{R}_{\mathbf{C}}(\mathbf{S} \odot (\mathbf{X} + t\mathbf{D}))$$
$$= t^2 \vec{\mathbf{D}}^T \tilde{\mathbf{S}}^T \mathbf{C}\tilde{\mathbf{S}}\vec{\mathbf{D}} + 2t\vec{\mathbf{X}}^T \tilde{\mathbf{S}}^T \mathbf{C}\tilde{\mathbf{S}}\vec{\mathbf{D}} + \vec{\mathbf{X}}^T \tilde{\mathbf{S}}^T \mathbf{C}\tilde{\mathbf{S}}\vec{\mathbf{X}},$$

where $\odot$ denotes the Hadamard product and $\tilde{\mathbf{S}} \in \mathbb{R}^{NL\times NL}$ the diagonal matrix with $\vec{\mathbf{S}}$ on its diagonal. Obviously, the restriction of f equals g where $\mathbf{C}$ is replaced by $\tilde{\mathbf{S}}\mathbf{C}\tilde{\mathbf{S}}$ which gives the first claim. The second claim follows since $\tilde{\mathbf{S}}\vec{\mathbf{D}} \in \pm\mathbb{R}_+^{NL}$, for $\mathbf{D} \in \pm\mathbb{R}_{\mathbf{S}}^{N\times L}$. ■

Proposition 6 states that if restricted to single orthants, $\mathcal{R}_{\mathbf{C}}(|\cdot|)$ behaves well along rays. In particular, the function $\mathcal{R}_{\mathbf{C}}(|\cdot|)$ is convex along all rays passing through the origin.

Corollary 7: For any $\mathbf{D} \in \mathbb{R}^{N\times L}$, the function

$$t \mapsto \mathcal{R}_{\mathbf{C}}(|t\mathbf{D}|)$$

is convex.

An important consequence of Proposition 6 is that if one had oracle knowledge on the orthant of $\mathbf{X}$, the program in (14) could be restricted accordingly and would become better conditioned. The naive approach thus would be to pick any initialization sharing the same sign with $\mathbf{X}$ and then restricting (14) to the corresponding orthant. Unfortunately, the possibly most common initialization, the back-projection of $\mathbf{Y}$ by the pseudo-inverse $\mathbf{A}^\dagger$ of $\mathbf{A}$

$$\mathbf{X}_0 = \mathbf{A}^\dagger\mathbf{Y} = \underset{\mathbf{Z}\in\mathbb{R}^{N\times L}}{\operatorname{argmin}} \|\mathbf{Z}\|_F, \quad \text{subject to } \mathbf{AZ} = \mathbf{Y}, \tag{16}$$

in general does not have this property as the following theorem shows.[1]

Theorem 8: For any (at least 2-sparse) vector $\mathbf{x}_0 \in \mathbb{R}^N$ (here $L = 1$), there exists $\mathbf{A} \in \mathbb{R}^{N\times N}$ and $\mathbf{y} \in \mathbb{R}^N$ such that $\mathbf{x}_0$ fulfills (16) and there exists a 2-sparse vector $\mathbf{x}^{\text{sp}}$ which solves the linear system, i.e., $\mathbf{Ax}^{\text{sp}} = \mathbf{y}$, but lies in another orthant.

Proof: Without loss of generality let $\mathbf{x}_0 \geq 0$ (entry-wise) and assume that its two first entries are non-zero. Define $\mathbf{x}^{\text{sp}} = (-a, b, 0, 0, \dots)$ with $a, b > 0$ and $\boldsymbol{\vartheta} = \mathbf{x}_0 - \mathbf{x}^{\text{sp}}$. Choose $a, b > 0$ such that

$$0 = \langle \boldsymbol{\vartheta}, \mathbf{x}_0\rangle = \|\mathbf{x}_0\|_2^2 + a(x_0)_1 - b(x_0)_2,$$

which is equivalent to

$$b = \frac{\|\mathbf{x}_0\|_2^2 + a(x_0)_1}{(x_0)_2}. \tag{17}$$

Now, let $\mathbf{A} \in \mathbb{R}^{N\times N}$ be any matrix with $\operatorname{span}(\boldsymbol{\vartheta}) = \ker(\mathbf{A})$ and define $\mathbf{y} = \mathbf{Ax}_0$. Then $\mathbf{x}_0$ is perpendicular to the kernel and thus the minimum norm solution of (16). Furthermore, $\mathbf{Ax}^{\text{sp}} = \mathbf{y}$ but $\mathbf{x}^{\text{sp}}$ lies in a different orthant than $\mathbf{x}_0$.

Instead of searching for suitable alternative initialization procedures, we propose to modify the optimization problem. Indeed, we can embed (2) into higher dimensional spaces and work with the augmented linear system

$$\mathbf{A}^{\pm}\mathbf{Z}^{\pm} = \mathbf{Y} \tag{18}$$

where $\mathbf{A}^{\pm} = (\mathbf{A}, -\mathbf{A}) \in \mathbb{R}^{m\times 2N}$ and $\mathbf{Z}^{\pm} \in \mathbb{R}^{2N\times L}$. Obviously, the original solution $\mathbf{X}$ solves (18) by defining

$$\mathbf{X}^{\pm} = \begin{pmatrix}\mathbf{X}\\ \mathbf{0}\end{pmatrix}. \tag{19}$$

More importantly, if we define $\mathbf{X}_+ \in \mathbb{R}_+^{N\times L}$ and $\mathbf{X}_- \in \mathbb{R}_+^{N\times L}$ as positive and negative part of $\mathbf{X}$ (containing only the positive/negative entries of $\mathbf{X}$ in absolute value and setting the rest to zero such that $\mathbf{X} = \mathbf{X}_+ - \mathbf{X}_-$), we have that the matrix

$$\mathbf{X}^{\pm}_{\text{pos}} = \begin{pmatrix}\mathbf{X}_+\\ \mathbf{X}_-\end{pmatrix}$$

solves (18), shares the structural complexity of $\mathbf{X}$, and lies within the positive orthant. To summarize the last lines, increasing the dimension of the linear system only by a factor of two, we can guarantee that a structured solution (from which the original $\mathbf{X}$ is directly recovered) may be found by applying an appropriate solver to (14) restricted to the positive orthant. From now on, we hence assume without loss of generality that $\mathbf{X}$ itself lies within the positive orthant and consider the restricted program

$$\min_{\mathbf{Z}\in\mathbb{R}_+^{N\times L}} \|\mathbf{AZ} - \mathbf{Y}\|_F^2 + \lambda\mathcal{R}_{\mathbf{C}}(\mathbf{Z}). \tag{20}$$

III. Optimization Via Gradient Descent

In order to approximate solutions to (20) we use gradient descent. For $F(\mathbf{Z}) := \|\mathbf{AZ} - \mathbf{Y}\|_F^2 + \lambda\mathcal{R}_{\mathbf{C}}(\mathbf{Z})$ the gradient is given by

$$\nabla F(\mathbf{Z}) = 2\mathbf{A}^T(\mathbf{AZ} - \mathbf{Y}) + 2\lambda\ \text{vec}^{-1}[\mathbf{C}\vec{\mathbf{Z}}],$$

where $\text{vec}^{-1}[\cdot] : \mathbb{R}^{NL} \to \mathbb{R}^{N\times L}$ inverts vectorization. To prevent gradient descent from leaving $\mathbb{R}_+^{N\times L}$, we replace the gradient by a projected version

$$[\tilde{\nabla}F(\mathbf{Z})]_{i,j} := \begin{cases} 0 & \text{if } [\nabla F(\mathbf{Z})]_{i,j} > 0,\ Z_{i,j} = 0,\\ [\nabla F(\mathbf{Z})]_{i,j} & \text{else.}\end{cases} \tag{21}$$

Note that $\tilde{\nabla}F(\mathbf{Z})$ still points into a descent direction. To compute a suitable step-size, we use the particular geometry of $\mathcal{R}_{\mathbf{C}}$. Let us define the descent ray function

$$f_{\mathbf{Z}}(t) = F(\mathbf{Z} - t\tilde{\nabla}F(\mathbf{Z})), \quad f_{\mathbf{Z}} : \mathbb{R} \to \mathbb{R},$$

for all $\mathbf{Z} \in \mathbb{R}^{N\times L}$. Note that $f_{\mathbf{Z}}$ can be written as

$$\begin{aligned} f_{\mathbf{Z}}(t) = &\left((\left\|\mathbf{A}\tilde{\nabla}F(\mathbf{Z})\right\|_F^2 + \lambda\text{vec}^{-1}\left[\overrightarrow{\tilde{\nabla}F(\mathbf{Z})}^T\mathbf{C}\,\overrightarrow{\tilde{\nabla}F(\mathbf{Z})}\right]\right)t^2\\ &+\Big(2\left\langle\mathbf{A}\tilde{\nabla}F(\mathbf{Z}), \mathbf{AZ} - \mathbf{Y}\right\rangle\\ &\quad -\lambda\vec{\mathbf{Z}}^T\mathbf{C}\,\overrightarrow{\tilde{\nabla}F(\mathbf{Z})} - \lambda\,\overrightarrow{\tilde{\nabla}F(\mathbf{Z})}^T\mathbf{C}\vec{\mathbf{Z}}\Big)t + c\\ = &\ at^2 + bt + c, \end{aligned}$$

where $c \in \mathbb{R}$ collects all terms not depending on t. We can now compute

$$\tilde{\sigma}_1 = \inf_{\substack{Z_{i,j}>0,\\ [\tilde{\nabla}F(\mathbf{Z})]_{i,j}>0}} \frac{Z_{i,j}}{\left|[\tilde{\nabla}F(\mathbf{Z})]_{i,j}\right|} \tag{22}$$

as the maximal allowed step-size to stay within $\mathbb{R}_+^{N\times L}$. Moreover, if $a > 0$, the function $f_{\mathbf{Z}}$ is strictly convex and the optimal (unconstrained) step-size is given by

$$\tilde{\sigma}_2 = -\frac{b}{2a}. \tag{23}$$

If $a \leq 0$, we set $\tilde{\sigma}_2 = 1$. We thus use

$$\sigma = \min\{\tilde{\sigma}_1, \tilde{\sigma}_2, 1\} \tag{24}$$

as step-size for our algorithm. Note that the choice of 1 as an upper bound for $\tilde{\sigma}_2$ and $\tilde{\sigma}_2$ is generic. Alternative choices would be $\frac{1}{\|\tilde{\nabla}F(\mathbf{Z})\|}$ or ∞.

[1] Although the initialization in (16) does not always provide the necessary orthant information, we mention that it often succeeds in numerical experiments.

Algorithm 1: Structural Sparse Recovery (SSR).

Given: $F(\mathbf{Z}) = \|\mathbf{AZ} - \mathbf{Y}\|_F^2 + \lambda \mathcal{R}_{\mathbf{C}}(\mathbf{Z})$

1: $\mathbf{X}_0 \leftarrow \mathbf{0}$
2: **while** stop condition is not satisfied **do**
3: $\quad \sigma_k \leftarrow \max\{\tilde{\sigma}_1, \tilde{\sigma}_2, 1\}$ ▷ see (22)–(24)
4: $\quad \mathbf{X}_{k+1} \leftarrow \mathbf{X}_k + \sigma_k \tilde{\nabla} F(\mathbf{X}_k)$ ▷ see (21)
5: **end while**
return $\mathbf{X}_{\text{rec}}$

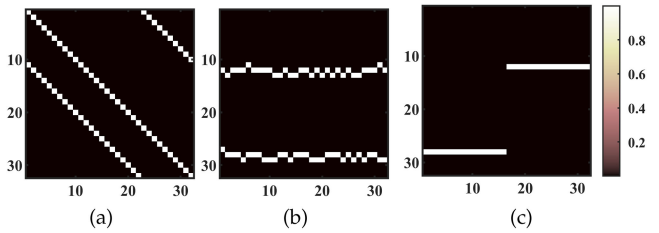


Fig. 2. Three examples of structures that may appear in applications: diagonals (a), oscillating lines (b), and partial row-sparsity (c).

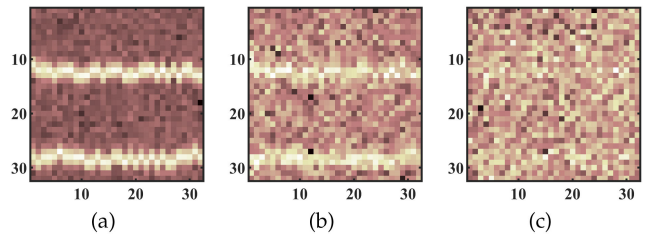


Fig. 3. Simulated data (oscillating lines) with low noise (a) (PSNR 24.92), medium noise (b) (PSNR 12.46), and high noise (c) (PSNR 1.97).

Remark 9: The structure encoding matrix $\mathbf{C}$ is a high-dimensional object. For an efficient implementation of Algorithm 1 it is crucial to construct $\mathbf{C}$ such that it allows fast vector-matrix multiplication. The heuristic we propose in Appendix A fulfills this requirement, see Remark 13.

The recent work [29] (which concentrates on $\mathbf{X}$ being a natural image) alternatively suggests to learn the projection operator onto the set of natural images via deep learning techniques and then to apply ADMM [10] to reconstruct $\mathbf{X}$ from linear measurements. In our setting this would correspond to learning the projection onto $\mathbb{S}_{\mathbf{C}}^s$ for a certain type of ground-truths, e.g., seismic data, and then applying ADMM. Note, however, that compared to ours this approach has two major drawbacks. First, training a deep network requires massive amounts of data that are not always available. Second, one cannot interpret the structure of the sparsity in $\mathbf{X}$ from a learned network whereas this is possible to some extent when using $\mathbf{C}$, cf. Fig. 8. It would be interesting to compare both approaches numerically in future work.

A. Convergence of Gradient Descent

As the objective functional is non-convex, the question arises under which assumptions Algorithm 1 can be expected to converge. The following observation provides a sufficient condition

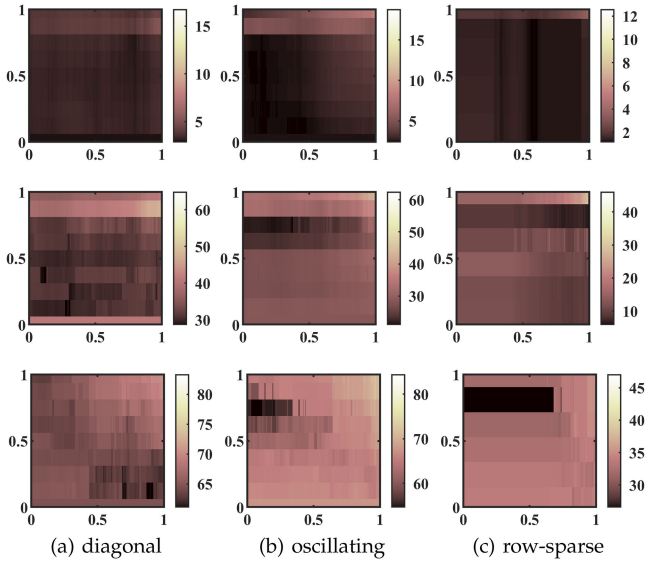


Fig. 4. Reconstruction error for different parameters α (y-axis), β (x-axis) and different noise levels low (top row), medium (middle row), and high (bottom row).

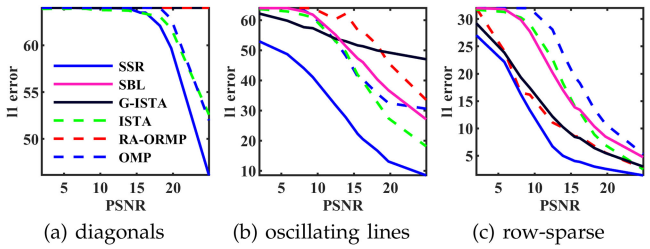


Fig. 5. Mean ℓ_1-error (scaled) vs. different noise levels for the different structures.

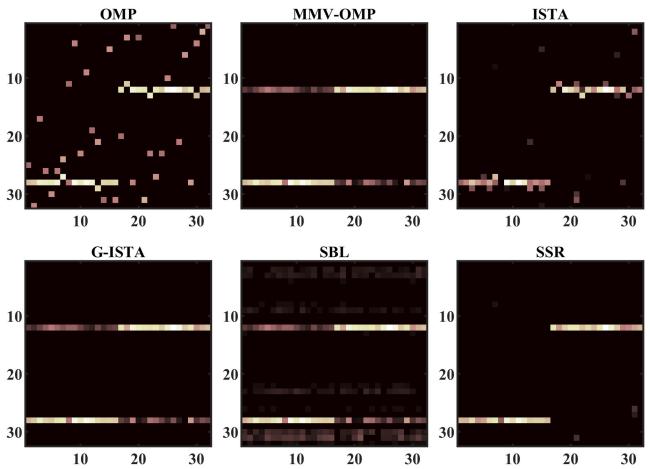


Fig. 6. Absolute value of reconstructions for different algorithms on row-sparse structures with medium noise level (PSNR 12.46).

for this to happen. We omit the proof which is straight-forward (recall that $\mathbf{C}$ has only non-negative entries).

Lemma 10: The objective function in (33) is bounded from below on $\mathbb{R}_+^{N\times L}$. If

$$\{\mathbf{Z} \in \mathbb{R}_+^{N\times L} : \mathcal{R}_{\mathbf{C}}(\mathbf{Z}) = 0\} \cap \ker(\mathbf{A}) = \{\mathbf{0}\}, \qquad (25)$$

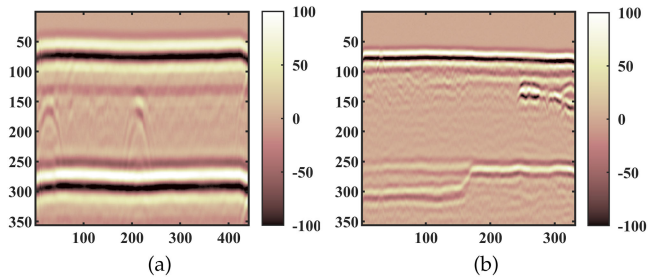


Fig. 7. Original ultrasonic non-destructive testing measurements of pores (a) and lack of fusion (b).

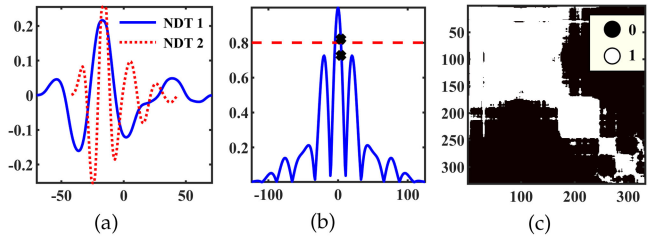


Fig. 8. Impulse functions (a), auto-correlation and the parameter α (dashed line, separates the fourth and fifth entry marked in black) (b) exemplary for NDT 1 in (a), and matrix $\mathbf{C}_\beta$ ($\beta = 0.75$) for the data in Fig. 7(b).

then it is coercive as well. Consequently, minimizers exist and all minimizers are contained in a finite ball around the origin.

Remark 11: The assumption in (25) is natural since it requires $\mathbf{A}$ to distinguish elementary structures from $\mathbf{0}$. When considering elementary structures generalizing row-sparsity, i.e., each column of the structured matrix contains at most one non-zero entry, the condition is fulfilled if $\mathbf{A}$ contains no zero columns. In case of more complex elementary structures with up to s-elements per columns in the structured matrix, (25) is implied by $\mathbf{A}$ having a *null-space property* of order s. In the context of solving underdetermined linear systems, null-space properties are a well-known concept from the compressed sensing literature [17].

IV. Numerical Experiments

Finally, let us demonstrate the power of the proposed Structural Sparse Recovery (SSR), see Algorithm 1. In Sections IV-A and IV-B, we perform basic tests on artificial data to validate our theoretical considerations. In Section IV-C, we present how SSR performs when applied to real-world data.

To quantify noise robustness, we use the notation of peak signal to noise ratio (PSNR). The reconstructed signal is denoted by $\mathbf{X}_{rec}$.

A. Parameter Heuristic Test

Let us first empirically validate whether the heuristic for constructing $\mathbf{C}$ as proposed in Appendix A yields practical results and whether the computed estimates for the therein used parameters α and β are meaningful. We choose three different $\mathbf{X} \in \mathbb{R}^{32\times 32}$, see Fig. 2, each of which contains two elementary structures (diagonal, slightly oscillating rows, partial rows), and a convolutional kernel $\mathbf{A} \in \mathbb{R}^{32\times 32}$ modeling the measurement process. We add three different levels of Gaussian noise – PSNR of 24.92 (low noise), 12.46 (moderate noise), and 1.97 (high noise) – to obtain corresponding measurements $\mathbf{Y} \in \mathbb{R}^{32\times 32}$, see Fig. 3. When constructing $\mathbf{C}$, for fixed $\mathbf{A}$ and $\mathbf{Y}$ there are only finitely many thresholds (α, β) of interest, cf. (26). We can thus reconstruct $\mathbf{X}$ from $\mathbf{A}$ and $\mathbf{Y}$ for all possible choices of α and β to obtain a benchmark for the heuristic. Note that we optimize in each reconstruction the additional parameters $\lambda_1, \lambda_2 \in [0, 1]$, cf. (33), over a fine grid.

Fig. 4 depicts the ℓ_1-error after scaling $\mathbf{X}_{\text{rec}}$ to have the same ℓ_1-norm as $\mathbf{X}$. This re-scaling balances the norm shrinkage caused by the penalty-term to allow for more meaningful comparisons. For the particularly chosen matrix $\mathbf{A}$, a choice of $\alpha = 0.7313$ considers two atoms that are shifted by at most one pixel as similar while for $\alpha = 0.99$ each atom is only similar to itself, i.e., we enforce row-sparsity. These thresholds can clearly be seen in Fig. 4 (in particular, when comparing a) and b) with c) in the low- and medium-noise cases). With increasing noise level it becomes more important to choose α sufficiently large in order to enforce structural sparsity. As the heuristic in Appendix A suggests, a choice of $\beta \leq \alpha$ yields optimal reconstruction results. As the noise level increases, the value of β has to be diminished such that more measurements are considered to be similar. This is as expected since more information is needed to reconstruct the signal in this case. Note that, for small to medium noise, the reconstruction performance is quite stable with respect to perturbations of the parameters α and β.

The diagonal case is most problematic in this regard (compare the color bars for the high-noise case) and demonstrates the limits of multiple measurements under this structure model. The strength of MMV is, to consider information of several columns at once to reconstruct the support. In the extreme case of row-sparsity, all columns can be taken into account at once. For localized structures such as Fig. 2(a) where the support changes fast the advantage of MMV is limited as only a small neighborhood carries information about the actual support of a column. Considering a larger neighborhood (i.e., decreasing β in our model) can help to improve the results. However, due to the fast changes of the support, we also need to consider more atoms as similar in the same step (i.e., decreasing α) which diminishes the gained information. Thus, if the noise level is too high, the model is not able to gain enough information for a suitable reconstruction. One way to overcome this problem might be to design a matrix $\mathbf{C}$ that is highly adapted to allow only diagonal structures.

B. Comparison

Let us now benchmark SSR against five state-of-the-art methods for the detection of sparsity and row-sparsity: Iterative Soft-Thresholding Algorithm (ISTA) [14], Group-Iterative Soft-Thresholding Algorithm (G-ISTA) [39], Orthogonal Matching Pursuit (OMP) and Rank Aware Order Recursive Matching Pursuit (RA-ORMP) [15], [33], and Sparse Bayesian Learning for row-sparsity (SBL) [35] (where we use the implementation of Z. Zhang [40]). ISTA is a technique for sparse recovery and G-ISTA its generalization to group sparse recovery of signals from underdetermined linear systems. RA-ORMP is an extension

TABLE I
SPARSITY OF THE STRUCTURES SHOWN IN FIG. 2 UNDER DIFFERENT NORMS

Structure Fig. 2	max column sparsity $\max_k \|\mathbf{x}_k\|_0$	row-sparsity $\|\mathbf{X}\|_{\text{row}-0}$	structural sparsity $\|\mathbf{X}\|_{\mathbf{C},0}$
a)	2	32	2
b)	2	6	2
c)	1	2	2

of the OMP algorithm for simultaneous sparse approximation. Last but not least, Sparse Bayesian Learning is a probabilistic regression method that has been developed in the context of machine learning.

We use for $\mathbf{X}$ the same structure types as considered in the previous section, see Fig. 2. Note that the different methods intend to minimize different sparsity measures. While OMP and ISTA minimize the individual sparsity of each column, G-ISTA, RA-ORMP and SBL minimize the row-sparsity. Hence, for different $\mathbf{X}$ some algorithms might be more suitable than others. Table I shows that Fig. 2(a) is clearly not row-sparse and thus favors our approach as well as the single measurement algorithms. In Fig. 2(b) row sparsity and structural sparsity are of comparable order. Finally, Fig. 2(c) has the same row and structural sparsity such that the comparison is not biased towards one single MMV method. When applying the different reconstruction algorithms we assume that the sparsity level is known, i.e., the iterative algorithms perform the exact amount of iterations needed. Moreover, the exact rank of X was given to RA-ORMP. Further hyper-parameters have been optimized over a grid to ensure best possible performance of all methods in the direct comparison.

To create our test data we use four different measurement matrices $\mathbf{A}$, two convolution matrices $\mathbf{A} \in \mathbb{R}^{32\times 32}$ with a Gauss kernel (similar to the previous section) as well as two random Gaussian matrices $\mathbf{A} \in \mathbb{R}^{10\times 32}$. This is a typical case of undersampling as considered in compressed sensing. In order to compare the results we re-scale the reconstruction $\mathbf{X}_{rec}$ as described above. In Fig. 5 the average ℓ_1-error of 300 runs is plotted for 11 different noise levels (PSNR of 2 to 25) where the range of the y-axis is restricted by the ℓ_1-norm of the ground-truth. It is clearly visible that the SSR algorithm outperforms all competitors, especially in the case of oscillating lines. This suggests that it is more robust to row-sparsity defects. A reconstruction example for each algorithm on the row-sparse structure is depicted in Fig. 6. We display the reconstructions in absolute value since small negative entries would otherwise change the colormap where 0 is supposed to be black, and thus make a comparison with the original more difficult.

C. Applications

Finally, we apply SSR to real-world data to show its capability of analyzing mixtures of different types of complex structures.

1) Non-Destructive Testing: The first example comes from the manufacturing industry in the field of non-destructive testing. Here, one tries to detect material defects and other anomalies from ultrasonic images of an object. In Fig. 7 ultrasonic images

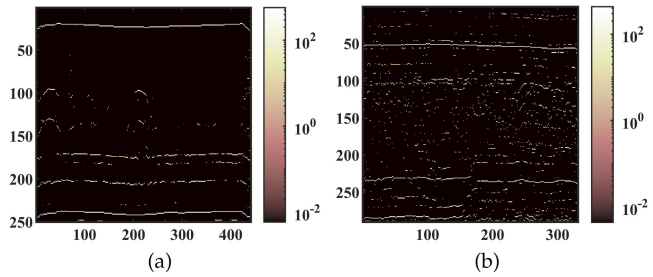


Fig. 9. Reconstructed structures (absolute value) in log-scale.

from scanning the weld seam of a steel pipe are depicted where the different structures result from the lateral signal and the back wall echo (horizontal lines) and the anomalies. The two images show different types of anomalies in the weld: in Fig. 7(a) we presumably see pores, whereas Fig. 7(b) shows a lack of fusion at the end of the pipe, where the last part of the weld seam has been ground. For representation of a received signal, one supposes that it can be obtained as a linear combination of time-shifted, energy-attenuated versions of the reconstructed pulse function (see Fig. 8(a)), where each shift is caused by an isolated flaw scattering the transmitted pulse [7]. This linear combination is described by the measurement matrix $\mathbf{A}$, a convolution matrix.

We expect a sparse reconstruction $\mathbf{X}_{rec}$ as we assume the material to have only few anomalies. Moreover, $\mathbf{X}_{rec}$ exhibits structure since adjacent measurements will give similar results. In particular, any anomaly will be seen in several adjacent measurements.

For the reconstruction of $\mathbf{X}$ we use the SSR model (33) with $\lambda_1 = 10^{-4}$ and $\lambda_2 = 10^6$. (We choose here $\lambda_1 = \mathcal{O}(L^{-2})$ following the intuition in Remark 14. Since the data is rather noisy we set $\lambda_2 = \mathcal{O}(L^2)$ instead of $\lambda_2 = \mathcal{O}(1)$ to prevent SSR from detecting multiple similar atoms per column.) The parameter α can be estimated using the auto-correlation of the wave impulse. Given the ultrasonic speed in the material, the time sampling of the signal, and the measurement setup, we can derive by how many pixels a signal from the same source can shift in between different measurements. For the examples given here a maximum shift of four respectively two pixels indicates a signal coming from the same source. Now, we choose α such that it separates the fourth and fifth respectively the second and third entry of the auto-correlation, cf. Fig. 8(b). For the example with the pores we set $\alpha = 0.8$ and for the example with the lack of fusion $\alpha = 0.78$. For the construction of $\mathbf{C}_2$ we choose a smaller value, namely $\alpha/5$, to enforce a gap between different structures. This way, the reconstruction results are further improved. As the noise level is low, we set $\beta = 0.75$ in both examples. We can see that for this choice the data structure already becomes evident in $\mathbf{C}_\beta$, e.g., the blocks appearing in Fig. 8(c) divide the measurements (columns in Fig. 8(c)) in three qualitatively different segments: the first segment contains two bands of which the lower one is diffuse, the second segment contains two well-separated bands, and the third contains the same two bands plus an additional artifact. The example thus perfectly fits into the structural sparsity framework of the paper. The reconstruction results are shown in Fig. 9. It can be seen that the method is able to detect both types of defects.

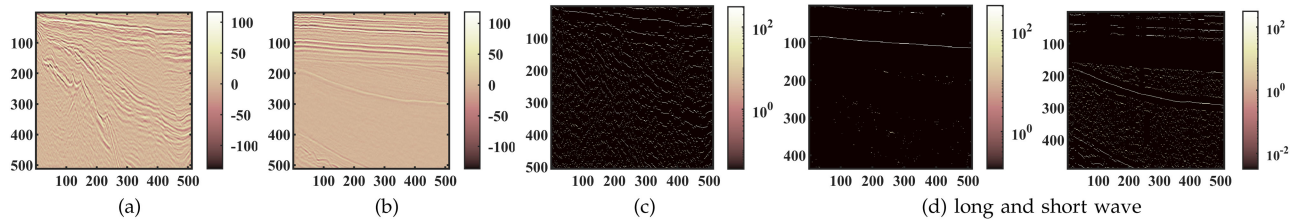


Fig. 10. Original seismic measurements (a), (b) and the reconstructed structures (absolute value) in log-scale (c), (d). Note that we show the reconstruction for the long and short wave impulse separately in (d).

TABLE II
PARAMETERS FOR SEISMIC EXPLORATION

	α	β	λ_1	λ_2
Example 1	0.5	0.4	10^{-6}	10^6
Example 2	0.7	0.6	1	$1.5 * 10^6$

2) Seismic Exploration: The second example is from the field of seismic exploration where the goal is to reconstruct soil layers in order to find natural resources such as gas and oil deposits or minerals. For this purpose, a wave is generated in a field experiment (e.g. by an explosion) and sensors that are arranged on a grid around the source register incoming seismic waves over time. The measured data form a 3D tensor of which we are looking at a slice (measured data along a straight line). For this reason, the example is very similar to the previous one and the reconstruction can be determined again with the same approach provided that the wave function is known (analogous to the pulse function in non-destructive testing). In [5], [6] it is described how to calculate a wave model as accurate as possible from the measurement data. This wave is then used again to set up the measurement matrix $\mathbf{A}$.

In Fig. 10 one can see slices from two different experiments with very different structures. While the seismic waves in the left image are all quite similar, the right image contains a mixture of longer and shorter waves. This indicates that for the right image the combination of a shorter and a longer wave impulse may be useful. A short wave impulse would reconstruct the diagonal structures in the middle and bottom left of the image well, but would not detect all horizontal structures in the upper part. This could only be achieved by using many short waves contradicting the assumption of sparsity. If we used only one longer wave instead, we would not capture the diagonal structures.

The parameters can be estimated as in the previous section. To compensate noise and model approximation we choose $\alpha \cdot 10^{-3}$ when constructing $\mathbf{C}_2$, i.e., no interference between structures is admitted. The values of the parameters are depicted in Table II. When choosing λ_1 and λ_2 we apply in Example 1 the same reasoning as in the non-destructive testing experiment; in Example 2, we increase both parameters to prevent long seismic waves from being replaced by multiple short ones.

The results are given in Fig. 10. Again, all kinds of structures have been reconstructed. The usage of two wave functions pays off as the row-like structures (Fig. 10(d), left) could be reconstructed separately from the diagonal-like structures.

3) Meteorology: As a third variation, we analyze the hourly precipitation in Germany from 25th to 28th of November 2008 (96 hours) using data of 932 weather stations shown in Fig. 11(a). Note that stations that were moved during this time or had too many missing values have not been taken into account. From these data we want to extract rain areas that are connected in time and space over as long a period as possible. We assume that the wind speed is constantly less than 75 km/h. Translated into our setting, this means that a connected rain area is a structure and $\mathbf{X} \in \mathbb{R}^{465\times 932}$ divides the total rain into disjoint, connected rain areas. The atoms of the system matrix $\mathbf{A}$ will represent rain showers of different lengths (two to six hours) observed at one station. Hence, $\mathbf{A} \in \{0, 1\}^{96\times 465}$ is the block band matrix

$$\mathbf{A} = [\mathbf{A}^2, \mathbf{A}^3, \ldots, \mathbf{A}^6] \text{ and } (A^k)_{j,l} = \begin{cases} 1 & 1 \leq j - l < k \\ 0 & \text{else} \end{cases}.$$

Instead of correlation between the atoms we use the starting time difference of these showers and instead of correlation between measurements we use the distance between the weather stations. This way, $\mathbf{C} \in \{0, 1\}^{932\cdot 465\times 932\cdot 465}$ encodes the geographical information. More precisely, $C_{(j,l),(j',l')} = 0$ only if the distance between the stations is smaller than 75 km/h times the starting time difference. Note that there is more meteorological data, such as wind direction or precise wind speeds of the according days available that could be used to construct a refined $\mathbf{C}$ such as wind direction or precise wind speeds of the according days. However, for this example we stick to the simpler model.

The measurement data shows the amount of precipitation in mm per hour. We simplify this to a binary matrix by setting $\mathbf{Y} \in \{0, 1\}^{96\times 932}$ where 0 indicates no rain and 1 indicates rain at a certain station at a certain time. This perfectly suits our system matrix $\mathbf{A}$ and guarantees that the linear system has a solution.

Despite the tremendous size of this system of equations it can be solved efficiently due to the fact that $\mathbf{C}$ is binary and can be factorized as a Kronecker product. Thus, the matrix multiplications can be reduced to summations. Moreover, we do not have to take the detour described in (18) as all involved matrices are non negative by definition.

In our computations we set $\lambda_1 = 200$ to just get one rain area per time interval and $\lambda_2 = 0$ since we do not enforce the structures to be sparse.

In Fig. 11 the two longest connected rain areas during the recorded period are depicted. In Fig. 11(d) we see a rain area moving from north to south-east starting at about 34 hours and ending at 51 and in Fig. 11(e) we see a rain area at the boarder to the Netherlands moving from south to north and starting at 56 hours and ending at about 75.

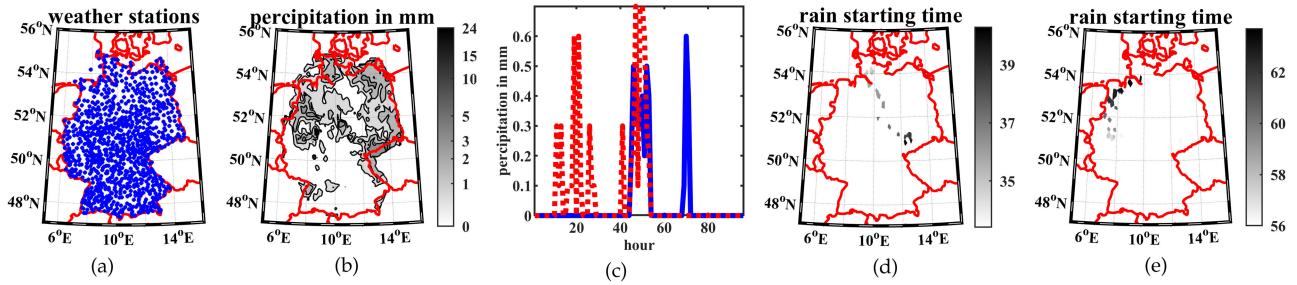


Fig. 11. Overall precipitation in Germany over 96 hours in November 2008 (b) (log scale) measured at locations (a). Exemplary data of two weather stations is given in (c). Tracked rain showers with longest duration (d), (e).

V. Conclusion

We presented in this paper a novel and user friendly model for handling structural sparsity in jointly sparse reconstruction tasks like MMV. We showed that our model is capable of representing various sparsity patterns of practical relevance. It, furthermore, lends itself to efficient reconstruction via projected gradient descent. In practice, the model parameters that determine the structures of interest can be heuristically learned from application specific environmental parameters, which are usually known to the practitioner. Comprehensive empirical studies confirmed efficacy of our method.

There remain several intriguing questions to be addressed in future work. First, we restricted the present work to introducing resp. motivating the model and empirically demonstrating its applicability. It would be desirable to derive supportive theoretical guarantees, both for the parameter heuristic in Appendix A and the (non-convex) projected gradient descent.

Second, it might be beneficial to modify the construction of $\mathbf{C}$ to allow entries with values in (0, 1), e.g., by using a soft-thresholding procedure. This could be used to enlargen the scope of representable structures, although it will presumably complicate parameter tuning.

Finally, it would be worthwhile to examine alternative constructions of $\mathbf{C}$ and compare their performance to our current heuristic in Appendix A.

Appendix A

A. Structural Sparsity Matrices

There is no unique admissible way to construct a structural matrix $\mathbf{C}$ for (6). For our purpose a simple heuristic approach suffices that constructs $\mathbf{C}$ from $\mathbf{A}$ and $\mathbf{Y}$. As our numerical examples in Section IV-C illustrate, the approach lends itself to further refinement depending on the concrete application. It is based on the idea that if two measurement vectors $\mathbf{y}_l = \mathbf{A}\mathbf{x}_l$ and $\mathbf{y}_{l'} = \mathbf{A}\mathbf{x}_{l'}$ are similar, then the corresponding atoms activated by $\mathbf{x}_l$ and $\mathbf{x}_{l'}$ in $\mathbf{A}$, i.e., $\{\mathbf{a}_j : j \in \operatorname{supp}(\mathbf{x}_l)\}$ and $\{\mathbf{a}_j : j \in \operatorname{supp}(\mathbf{x}_{l'})\}$, should be similar as well.

Definition 12: Let $0 < \alpha, \beta < 1$ be two predefined threshold parameters. We say that two atoms $\mathbf{a}_j$ and $\mathbf{a}_{j'}$ are *similar* if

$$\operatorname{Corr}(\mathbf{a}_j, \mathbf{a}_{j'}) = \frac{|\langle \mathbf{a}_j, \mathbf{a}_{j'} \rangle|}{\|\mathbf{a}_j\|_2 \|\mathbf{a}_{j'}\|_2} > \alpha.$$

Two measurements $\mathbf{y}_l$ and $\mathbf{y}_{l'}$ are *similar* if

$$\operatorname{Corr}(\mathbf{y}_l, \mathbf{y}_{l'}) = \frac{|\langle \mathbf{y}_l, \mathbf{y}_{l'} \rangle|}{\|\mathbf{y}_l\|_2 \|\mathbf{y}_{l'}\|_2} > \beta.$$

Let $\bar{\mathbf{A}}$ and $\bar{\mathbf{Y}}$ denote copies of $\mathbf{A}$ and $\mathbf{Y}$ with re-normalized columns, i.e., $\bar{\mathbf{A}}^T\bar{\mathbf{A}}$ and $\bar{\mathbf{Y}}^T\bar{\mathbf{Y}}$ contain the pairwise correlation of all atoms and measurement vectors. We define

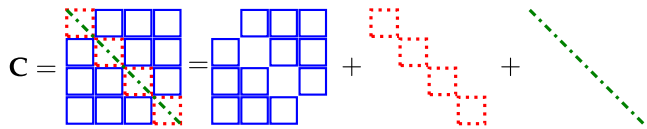
$$\mathbf{C}_\beta \in \{0, 1\}^{L \times L}, \qquad (C_\beta)_{j,k} = \begin{cases} 1 & (\bar{Y}^T\bar{Y})_{j,k} > \beta \\ 0 & (\bar{Y}^T\bar{Y})_{j,k} \leq \beta \end{cases}, \tag{26}$$

and $\mathbf{C}_\alpha \in \mathbb{R}^{N \times N}$ accordingly where $\bar{\mathbf{Y}}$ is replaced with $\bar{\mathbf{A}}$. Consequently, $\mathbf{C}_\beta$ is 1 for each pair of similar measurement vectors and 0 otherwise. A naive design for $\mathbf{C}$ is then

$$\mathbf{C}_{\text{simple}} = \mathbf{C}_\beta \otimes (\mathbb{1} - \mathbf{C}_\alpha) \in \mathbb{R}^{NL \times NL}. \tag{27}$$

Now $C_{(j,l),(j',l')}$ is 1 whenever the measurement vectors $\mathbf{y}_l$ and $\mathbf{y}_{l'}$ are similar but the chosen atoms $\mathbf{a}_j$ and $\mathbf{a}_{j'}$ are not.

Constructing $\mathbf{C}$ as in (27), (5) penalizes whenever two measurement vectors $\mathbf{y}_l$ and $\mathbf{y}_{l'}$ are similar but the atoms of $\mathbf{A}$ activated by $\mathbf{x}_l$ and $\mathbf{x}_{l'}$ are not. It does not penalize activation of similar atoms $\mathbf{a}_j$ and $\mathbf{a}_{j'}$ by a single $\mathbf{x}_l$, i.e., in the same measurement vector $\mathbf{y}_l$. Numerical experiments show that blurring effects in the reconstruction are a consequence. To overcome the problem, we instead construct $\mathbf{C}$ as a composition of L^2 blocks of size $N \times N$ in the following way:



$$= \mathbf{C}_1 + \mathbf{C}_2 + \mathbf{C}_3.$$

Here, $\mathbf{C}_1$ is the penalization for pairwise different measurement vectors, $\mathbf{C}_2$ compares each column of the solution with itself and $\mathbf{C}_3$ compares each element with itself. Since the comparison of each element with itself does not give any structural information, we set $\mathbf{C}_3 = \mathbf{0}$. Pairwise different measurements can be treated as in the example above and thus

$$\mathbf{C}_1 = (\mathbf{C}_\beta - \mathbf{Id}) \otimes (\mathbb{1} - \mathbf{C}_\alpha), \tag{28}$$

where using $(\mathbf{C}_\beta - \mathbf{Id})$ instead of $\mathbf{C}_\beta$ sets $\mathbf{C}_1$ to zero on the diagonal blocks. For $\mathbf{C}_2$ we use the contrary heuristic: whenever two elements in the same column $\mathbf{x}_l$ of $\mathbf{X}$ are non-zero, the corresponding activated atoms in $\mathbf{A}$ should not be similar as otherwise one of them would be redundant. We set

$$\mathbf{C}_2 = \mathbf{Id} \otimes (\mathbf{C}_\alpha - \mathbf{Id}), \tag{29}$$

where subtracting the identity sets $\mathbf{C}_2$ to zero on the main diagonal. This construction of $\mathbf{C}$ exhibits sound performance in numerical experiments, see Section IV.

Remark 13: The Kronecker form of the above matrices allows fast matrix-vector multiplication. This becomes important later on as $\mathbf{C}$ is of high dimension and multiplication by $\mathbf{C}$ is a fundamental operation for our numerical simulations.

Choosing α and β: When using the above construction method, we need to choose suitable parameters α and β controlling the shape of $\mathbf{C}$. As in Proposition 3, we assume here that each column of $\mathbf{X}$ contains at most one non-zero entry per active elementary structure, cf. Remark 4.

A meaningful parameter α can be directly deduced from prior knowledge on the concrete application. See Section IV for examples on how these priors can look like and how α is derived from them. If α is given, β can be estimated from $\mathbf{Y}$ and $\mathbf{A}$. To this end, let $\mathbf{y}, \mathbf{y}'$ be two columns of $\mathbf{Y}$ and $\mathbf{x}, \mathbf{x}'$ the corresponding columns of $\mathbf{X}$, i.e., $\mathbf{y} = \mathbf{A}\mathbf{x}$ and $\mathbf{y}' = \mathbf{A}\mathbf{x}'$. Recall from (27) that at this point we are interested in determining β such that, by Definition 12, $\mathbf{y}$ and $\mathbf{y}'$ are similar only if $\mathbf{x}$ and $\mathbf{x}'$ activate similar atoms. For simplicity, we assume here that $\mathbf{A}$ has unit norm columns; the argument can be easily adapted to the general case. Assume s elementary structures are active in $\mathbf{X}$ where we do not need to know s. We introduce copies $\boldsymbol{\xi}, \boldsymbol{\xi}' \in \mathbb{R}^s$ of $\mathbf{x}, \mathbf{x}'$ which are reduced to the support of $\mathbf{x}, \mathbf{x}'$ and re-ordered such that ξ_k and ξ'_k belong to the same elementary structure, for all $k \in [s]$. If there is no noise on the measurements, we can write $\mathbf{y}$ and $\mathbf{y}'$ as

$$\mathbf{y} = \sum_{k=1}^{s} \xi_k \mathbf{a}_{j_k} \quad \text{and} \quad \mathbf{y}' = \sum_{k=1}^{s} \xi'_k \mathbf{a}_{j'_k},$$

where $\mathbf{a}_{j_k}, \mathbf{a}_{j'_k}$ are the atoms activated by the k-th structure in $\mathbf{x}$ resp. $\mathbf{x}'$. Hence, the inner product of $\mathbf{y}$ and $\mathbf{y}'$ is

$$\langle \mathbf{y}, \mathbf{y}' \rangle = \sum_{k=1}^{s} \xi_k \xi'_k \langle \mathbf{a}_{j_k}, \mathbf{a}_{j'_k} \rangle + \sum_{\substack{k,k'=1 \\ k \neq k'}}^{s} \xi_k \xi'_{k'} \langle \mathbf{a}_{j_k}, \mathbf{a}_{j'_{k'}} \rangle. \tag{30}$$

The first term represents the correlation of each of the s structures with itself, while the second represents interference between different structures. Note that the first sum only contains inner products of atoms $\mathbf{a}_{j_k}$ and $\mathbf{a}_{j'_k}$ activated by the same structure. Since we consider here the situation that activated atoms are similar, by Definition 12, the absolute value of these inner products is limited from below by α. If we assume that

A) there are no sudden phase shifts in the elementary structures of our ground-truth $\mathbf{X}$,
i.e., $\operatorname{sign}(\xi_k) = \operatorname{sign}(\xi'_k)$, for all $k \in [s]$,

then all terms in the first sum are positive. If, in addition,[2]

B) the interference between elementary structures is bounded by $0 < \varepsilon \ll \alpha \langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle$,

we know from (30) that

$$\langle \mathbf{y}, \mathbf{y}' \rangle \geq \alpha \langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle - \varepsilon \quad \text{and} \quad \|\mathbf{y}\|^2 \leq \|\boldsymbol{\xi}\|_2^2 + \varepsilon. \tag{31}$$

[2]Both assumptions, (A) and (B), are mild and often hold in applications. Assumption (B), e.g., holds whenever the elementary structures are separated or destructive interference is occurring.

For the correlation of the two measurements we thus obtain

$$\begin{aligned} \operatorname{Corr}(\mathbf{y}, \mathbf{y}') = \frac{\langle \mathbf{y}, \mathbf{y}' \rangle}{\|\mathbf{y}\|_2 \|\mathbf{y}'\|_2} &\geq \frac{\alpha \langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle - \varepsilon}{\sqrt{\|\boldsymbol{\xi}\|_2^2 + \varepsilon} \sqrt{\|\boldsymbol{\xi}'\|_2^2 + \varepsilon}} \\ &\approx \alpha \operatorname{Corr}(\boldsymbol{\xi}, \boldsymbol{\xi}'), \end{aligned} \tag{32}$$

such that we can choose $\beta = \alpha \operatorname{Corr}(\boldsymbol{\xi}, \boldsymbol{\xi}')$. Note that $\operatorname{Corr}(\boldsymbol{\xi}, \boldsymbol{\xi}')$ is close to 1 whenever the entries of $\mathbf{X}$ do hardly vary along elementary structures, and it is small for highly fluctuating entries along elementary structures. In applications the (expected) correlation $\operatorname{Corr}(\boldsymbol{\xi}, \boldsymbol{\xi}')$ should be given approximately. Let us mention that the above assumptions are not required to hold for arbitrary $\mathbf{y}, \mathbf{y}'$ and elementary structures. To have a reliable heuristic parameter choice, it suffices if they apply to the majority of the measurements.

In the presence of noise, we decrease the estimate (32) relative to the noise level. By this the measurements are still considered similar even if they are corrupted by noise. The simulations in Section IV show efficacy of the heuristic.

Let us finally mention that while testing in numerical experiments the performance of (20) with $\mathbf{C}$ as constructed above, it turned out that depending on the concrete problem setting the tuning of λ is challenging. The main problem appears to be the different contributions of $\mathbf{C}_1$ and $\mathbf{C}_2$ in (28) and (29) to the regularizing function $\mathcal{R}_{\mathbf{C}}$. While $\mathbf{C}_1$ compares pairwise different measurements and enforces global structure, $\mathbf{C}_2$ compares each column of the solution with itself enforcing sparsity along the columns. Consequently, it is beneficial to additionally balance between $\mathbf{C}_1$ and $\mathbf{C}_2$ by introducing parameters $\lambda_1, \lambda_2 > 0$ and defining $\mathbf{C}_\lambda = \lambda_1 \mathbf{C}_1 + \lambda_2 \mathbf{C}_2$, for $\lambda = (\lambda_1, \lambda_2)$. Then, the program (20) becomes

$$\min_{\mathbf{Z} \in \mathbb{R}_+^{N \times L}} \|\mathbf{A}\mathbf{Z} - \mathbf{Y}\|_F^2 + \lambda_1 \mathcal{R}_{\mathbf{C}_1}(\mathbf{Z}) + \lambda_2 \mathcal{R}_{\mathbf{C}_2}(\mathbf{Z}). \tag{33}$$

The preceding results — Lemma 5, Propostion 6, and Corollary 7 — extend in a straight-forward way to (33) by using linearity of $\mathcal{R}$ in $\mathbf{C}$, i.e., $\lambda_1 \mathcal{R}_{\mathbf{C}_1}(\mathbf{Z}) + \lambda_2 \mathcal{R}_{\mathbf{C}_2}(\mathbf{Z}) = \mathcal{R}_{\mathbf{C}_\lambda}(\mathbf{Z})$.

Remark 14: Let us briefly comment on why λ_1 and λ_2 may notably differ in applications. Recall the proof of Proposition 3. In particular, the bounds in inequality (10) are dominated by the influence of $\mathbf{C}_1$. Considering both matrix parts separately, one could refine the bound to be

$$1 \leq \vec{\mathbf{X}_k}^T \mathbf{C}_1 \vec{\mathbf{X}_j} \leq L^2 \quad \text{and} \quad 0 \leq \vec{\mathbf{X}_k}^T \mathbf{C}_2 \vec{\mathbf{X}_j} \leq L.$$

The second term penalizes structures that are overlapping or close to each other. Since this hardly happens throughout all measurements, we expect the upper bound to be overpessimistic. In applications, one would rather have $\mathcal{O}(1)$ than $\mathcal{O}(L)$. Considering $\mathbf{C}_1$, i.e., the first inequality, however, a scaling of L^2 seems realistic whenever the structures are active throughout all measurements. Hence, the influence of $\mathbf{C}_1$ on the penalty term is up to L^2-times larger compared to $\mathbf{C}_2$ which can be compensated by scaling λ_2 accordingly. In addition, λ_1 balances the structural sparsity and must be chosen smaller the more structures appear in the data while λ_2 handles the sparsity within single structures and is mostly independent of the number of active elementary structures. As Section IV shows, a parameter setup with very small $\lambda_1 = \mathcal{O}(L^{-2})$ and $\lambda_2 \gg 1$ is not unusual.

ACKNOWLEDGMENT

The figures in Section IV-C3 were made using M Map, a mapping package for MATLAB by R. Pawlowicz ([Online]. Available: https://www.eoas.ubc.ca/rich/map.html).

REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[2] D. Angelosante, G. B. Giannakis, and E. Grossi, "Compressed sensing of time-varying signals," in *Proc. 16th Int. Conf. Digit. Signal Process.*, 2009, pp. 1–8.

[3] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.

[4] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, "Distributed compressed sensing," Tech. Rep. ECE-0612, Dept. Elect. Comput. Eng., Rice University, Dec. 2006.

[5] F. Bossmann and J. Ma, "Asymmetric chirplet transform for sparse representation of seismic data," *Geophysics*, vol. 80, no. 6, pp. WD89–WD100, 2015.

[6] F. Bossmann and J. Ma, "Asymmetric chirplet transform-part 2: Phase, frequency, and chirp rate," *Geophysics*, vol. 81, no. 6, pp. V425–V439, 2016.

[7] F. Boßmann, G. Plonka, T. Peter, O. Nemitz, and T. Schmitte, "Sparse deconvolution methods for ultrasonic NDT," *J. Nondestruct. Eval.*, vol. 31, no. 3, pp. 225–244, 2012.

[8] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[9] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[10] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[11] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex algebraic geometry of linear inverse problems," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput.*, 2010, pp. 699–703.

[12] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, Dec. 2006.

[13] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.

[14] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[15] M. Davies and Y. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1135–1146, Feb. 2012.

[16] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[17] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Basel, Switzerland, Birkhäuser Basel, 2013.

[18] P. Heins, M. Moeller, and M. Burger, "Locally sparse reconstruction using the $\ell^{1,\infty}$-norm," *Inverse Problems Imag.*, vol. 9, no. 4, pp. 1093–1137, 2015.

[19] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *J. Mach. Learn. Res.*, vol. 12, no. 11, pp. 3371–3412, 2011.

[20] J. Huang and T. Zhang, "The benefit of group sparsity," *Ann. Statist.*, vol. 38, no. 4, pp. 1978–2004, 2010.

[21] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 35, no. 3, pp. 400–401, 1987.

[22] S. Lee, Y. Liao, M. Seo, and Y. Shin, "Oracle Estimation of a Change Point in High-Dimensional Quantile Regression," . *Amer. Statist. Assoc.*, vol. 113, no. 523, pp. 1184–1194, 2018, *arXiv:1411.3062*.

[23] L. Lian, A. Liu, and V. Lau, "Exploiting dynamic sparsity for downlink FDD-massive MIMO channel tracking," *IEEE Trans. Signal Process.*, vol. 67, no. 8, pp. 2007–2021, Apr. 2019.

[24] X. Lu, H. Yuan, P. Yan, Y. Yuan, and X. Li, "Geometry constrained sparse coding for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1648–1655.

[25] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, 2007.

[26] T. Peleg, Y. C. Eldar, and M. Elad, "Exploiting statistical dependencies in sparse representations for signal recovery," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2286–2303, May 2012.

[27] B. D. Rao, "Signal processing with the sparseness constraint," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seattle, Washington, USA, 1998, pp. 1861–1864.

[28] X. Rao and V. K. N. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, Jun. 2014.

[29] J. Rick Chang, C.-L. Li, B. Poczos, B. Vijaya Kumar, and A. C. Sankaranarayanan, "One network to solve them all-solving linear inverse problems using deep projection models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5888–5897.

[30] E. Y. Sidky and X. Pan, "Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization," *Phys. Med. Biol.*, vol. 53, no. 17, pp. 4777–4807, 2008.

[31] J.-L. Starck, F. Murtagh, and J. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[32] J. A. Tropp, "Algorithms for simultaneous sparse approximation: Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, 2006.

[33] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.

[34] N. Vaswani and W. Lu, "Modified-CS: Modifying compressive sensing for problems with partially known support," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4595–4607, Sep. 2010.

[35] D. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3704–3716, Jul. 2007.

[36] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[37] Y. Wu *et al.*, "Accelerated MR diffusion tensor imaging using distributed compressed sensing," *Magn. Reson. Med.*, vol. 71, no. 2, pp. 764–772, 2014.

[38] L. Yu, H. Sun, J.-P. Barbot, and G. Zheng, "Bayesian compressive sensing for cluster structured sparse signals," *Signal Process.*, vol. 92, no. 1, pp. 259–269, 2012.

[39] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Society: Ser. B (Stat. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.

[40] Z. Zhang, "(5) SBL (sparse Bayesian learning)," ver. 1.1 (02/12/2011), Accessed: Feb. 9, 2021. [Online]. Available: http://dsp.ucsd.edu/ zhilin/ Software.html

[41] Z. Zhang and B. D. Rao, "Sparse signal recovery in the presence of correlated multiple measurement vectors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Sheraton Dallas Hotel, Dallas, Texas, USA, 2010, pp. 3966–3989.

[42] Z. Zhang and B. D. Rao, "Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 2009–2015, Apr. 2013.

[43] H. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.

[44] H. Zhu, G. Leus, and G. B. Giannakis, "Sparsity-cognizant total least-squares for perturbed compressive sampling," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2002–2016, May 2011.

[45] J. Ziniel and P. Schniter, "Dynamic compressive sensing of time-varying signals via approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5270–5284, Nov. 2013.

[46] J. Ziniel and P. Schniter, "Efficient high-dimensional inference in the multiple measurement vector problem," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 340–354, Jan. 2013.